Linked Data in Production

Moving Beyond Ontologies

David Newbury
Assistant Director, Software and User Experience, Getty

CNI Spring Project Briefings, March 26, 2024

Getty

Thank you for inviting me to present today.

I am a **technologist** working in the cultural heritage sector.

I lead the **public digital** team, developing applications to support Getty's mission.

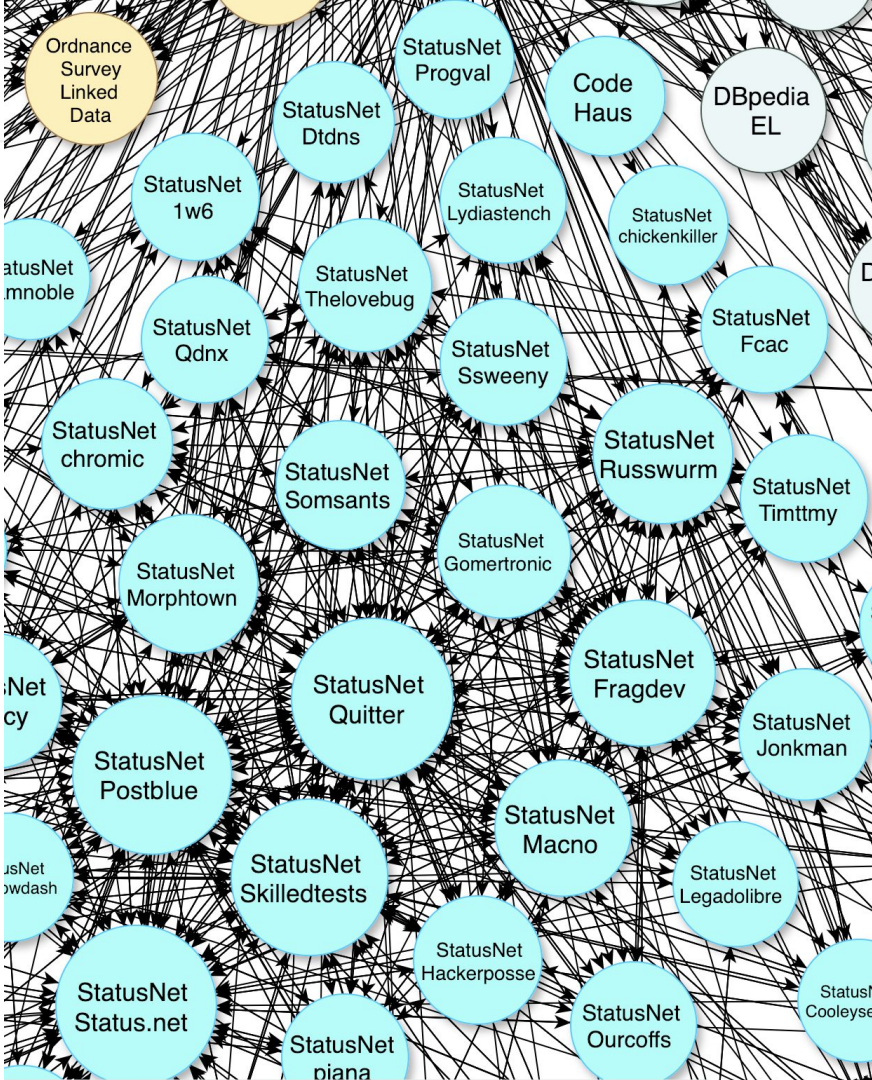
I work for **Getty** in Los Angeles, which is a library/archive/museum/research center.

One of our major areas of digital leadership is in the use of **Linked Data** for **cultural heritage**.

Getty

Linked Open Data is a set of technologies that attempt to translate some of the best practices of the Web for use with structured data:

- The use of URLs as identifiers
- Networks of information, not tables
- Formal, shared standards for description



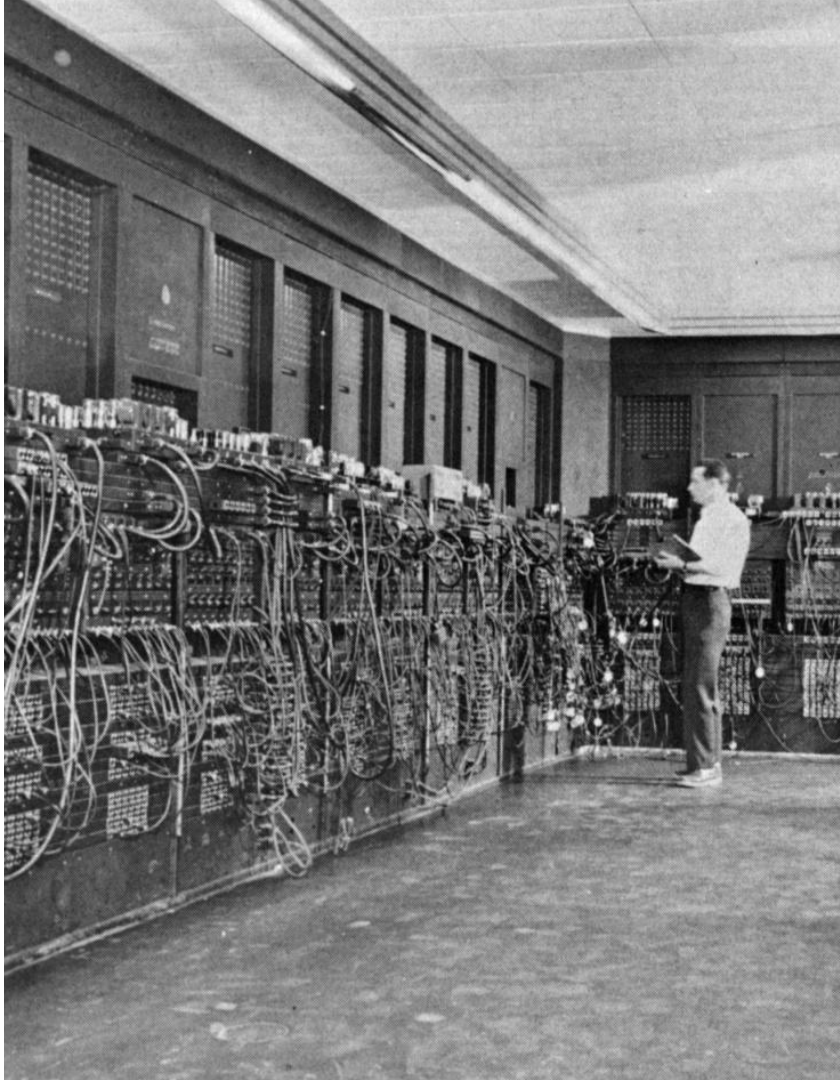
Getty

As it turns out, Linked Data is not wildly successful.

But we all still talk about it a lot.

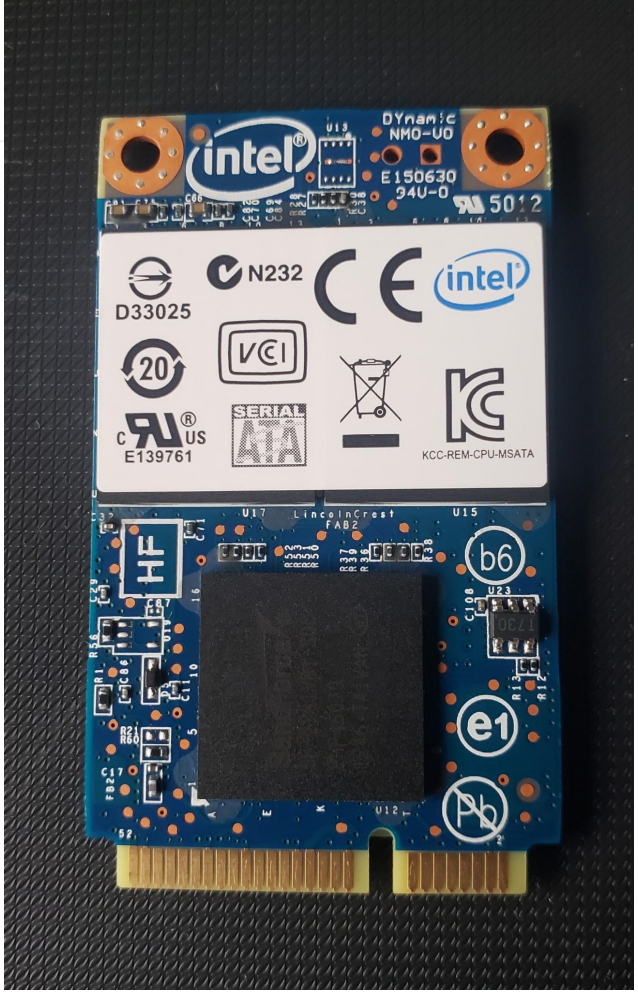
Linked Data's appeal in cultural heritage is a technological solution to a social problem:

Cheap storage, ubiquitous connectivity, and search algorithms **recontextualize the labor** behind cultural heritage data work.



Mass digitization, computational metadata generation, and decades of catalouging mean that our institutions have more data to provide than we have the **ability to provide context** for.

Data overload and limited user attention are the collections access problems of the next decade.



Our always-connected culture means that our collections are increasingly seen as part of a **single digital ecosystem**—

And the questions that are being asked require information that **extends beyond** the boundaries of any one institution.

WIKIPEDIA
The Free Encyclopedia

Create account Log in

List of museums in the United States

4 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

A list of museums in the United States by state. According to a government statement, there are more than 35,000 museums in the US.[1]

Alabama [edit]
See List of museums in Alabama.
See also Category:Museums in Alabama.

Alaska [edit]
See List of museums in Alaska.
See also Category:Museums in Alaska.

Arizona [edit]
See List of museums in Arizona.
See also Category:Museums in Arizona.

Arkansas [edit]
See List of museums in Arkansas.
See also Category:Museums in Arkansas.

California [edit]
See List of museums in California.
See also Category:Museums in California.

Colorado [edit]
See List of museums in Colorado.
See also Category:Museums in Colorado.

Connecticut [edit]
See List of museums in Connecticut.
See also Category:Museums in Connecticut.

Delaware [edit]
See List of museums in Delaware.
See also Category:Museums in Delaware.

Florida [edit]
See List of museums in Florida.
See also Category:Museums in Florida.

Georgia [edit]
See List of museums in Georgia.
See also Category:Museums in Georgia (U.S. state).

Hawaii [edit]
See List of museums in Hawaii.
See also Category:Museums in Hawaii.

Idaho [edit]
See List of museums in Idaho.
See also Category:Museums in Idaho.

Illinois [edit]
See List of museums in Illinois.
See also Category:Museums in Illinois.
See also List of museums and cultural institutions

Montana [edit]
See List of museums in Montana.
See also Category:Museums in Montana.

Nebraska [edit]
See List of museums in Nebraska.
See also Category:Museums in Nebraska.

Nevada [edit]
See List of museums in Nevada.
See also Category:Museums in Nevada.

New Hampshire [edit]
See List of museums in New Hampshire.
See also Category:Museums in New Hampshire.

New Jersey [edit]
See List of museums in New Jersey.
See also Category:Museums in New Jersey.

New Mexico [edit]
See List of museums in New Mexico.
See also Category:Museums in New Mexico.

New York [edit]
See List of museums in New York.
See also Category:Museums in New York (state).
See also List of university art museums and galleries in New York State.
See also List of museums and cultural institutions in New York City.
See also List of museums in New York City.
See also List of museums on Long Island.

North Carolina [edit]
See List of museums in North Carolina.
See also Category:Museums in North Carolina.

North Dakota [edit]
See List of museums in North Dakota.
See also Category:Museums in North Dakota.

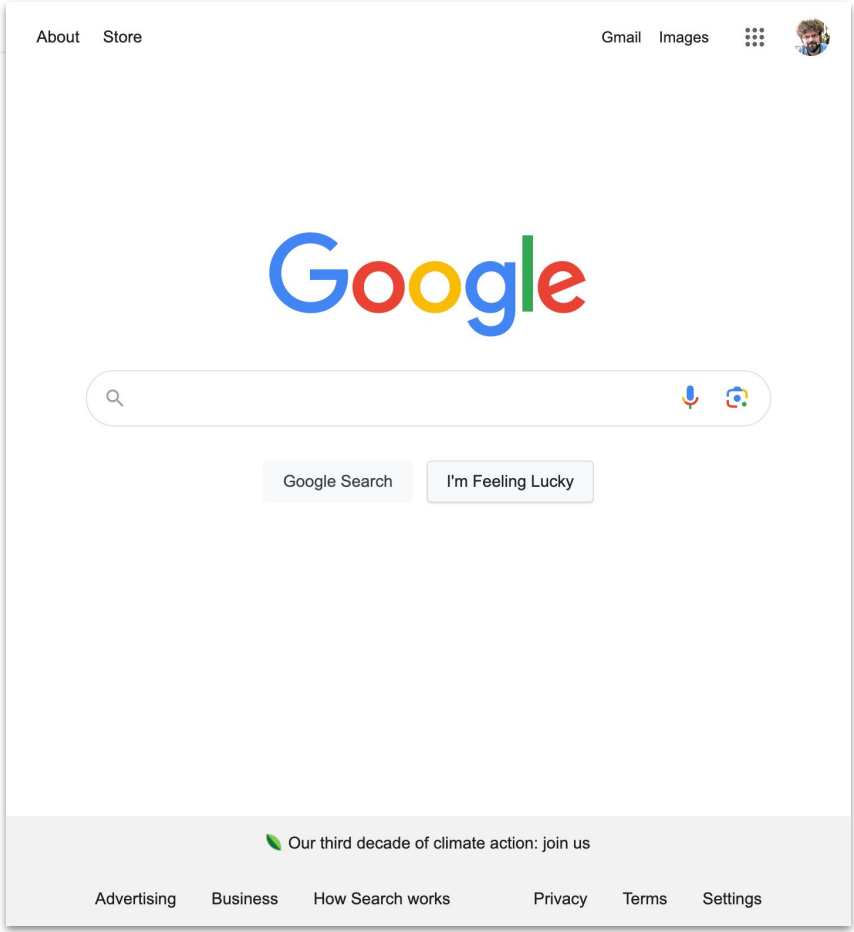
Ohio [edit]
See List of museums in Ohio.
See also Category:Museums in Ohio.

Oklahoma [edit]
See also List of museums in Oklahoma.
See also Category:Museums in Oklahoma.

Oregon [edit]
See List of museums in Oregon.
See also Category:Museums in Oregon.

And commercial tools have given users the expectation that information is available for the asking—

Eliding the labor and capital needed to create, curate, and maintain that information.



Linked Data has been seen as a solution:

It provide structures that manage the **scale of data** we create,

identifiers that maintain **authority** in a globally distributed environment,

and ontologies that enable complex **data retrieval** across datasets.

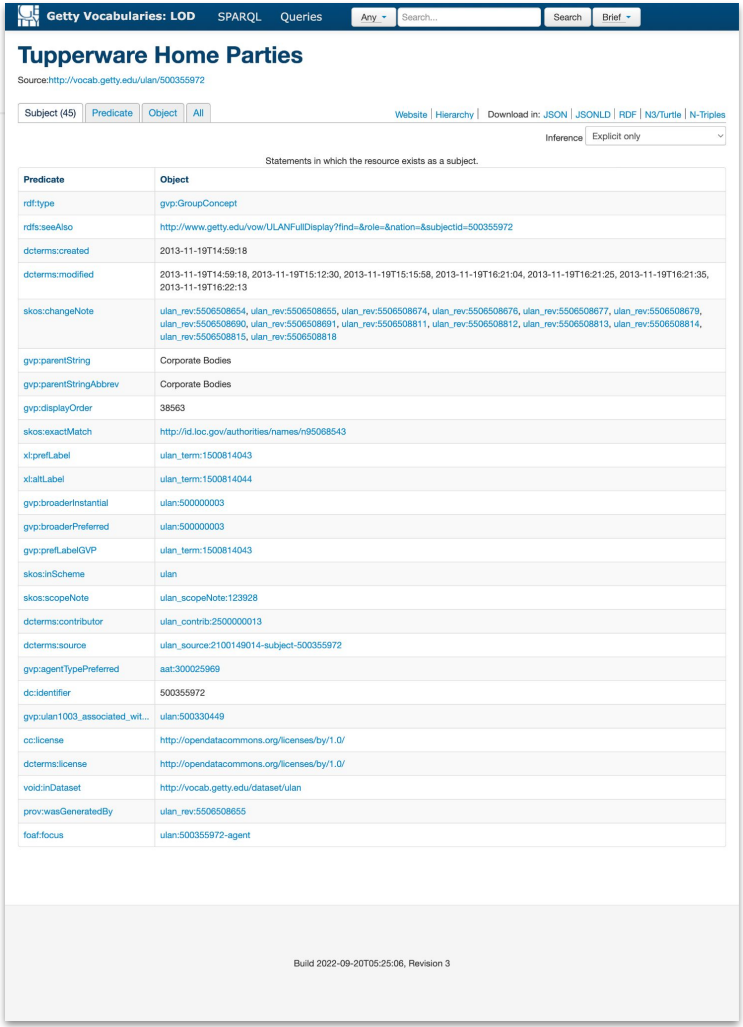
What came before:

Laying the Groundwork

Getty

In 2014, the **Getty Vocabularies** were launched as Linked Data.

This, alongside the work at Yale Center for British Art, Rijksmuseum, and the British Museum, demonstrated the **feasibility** of LOD within the museum community.



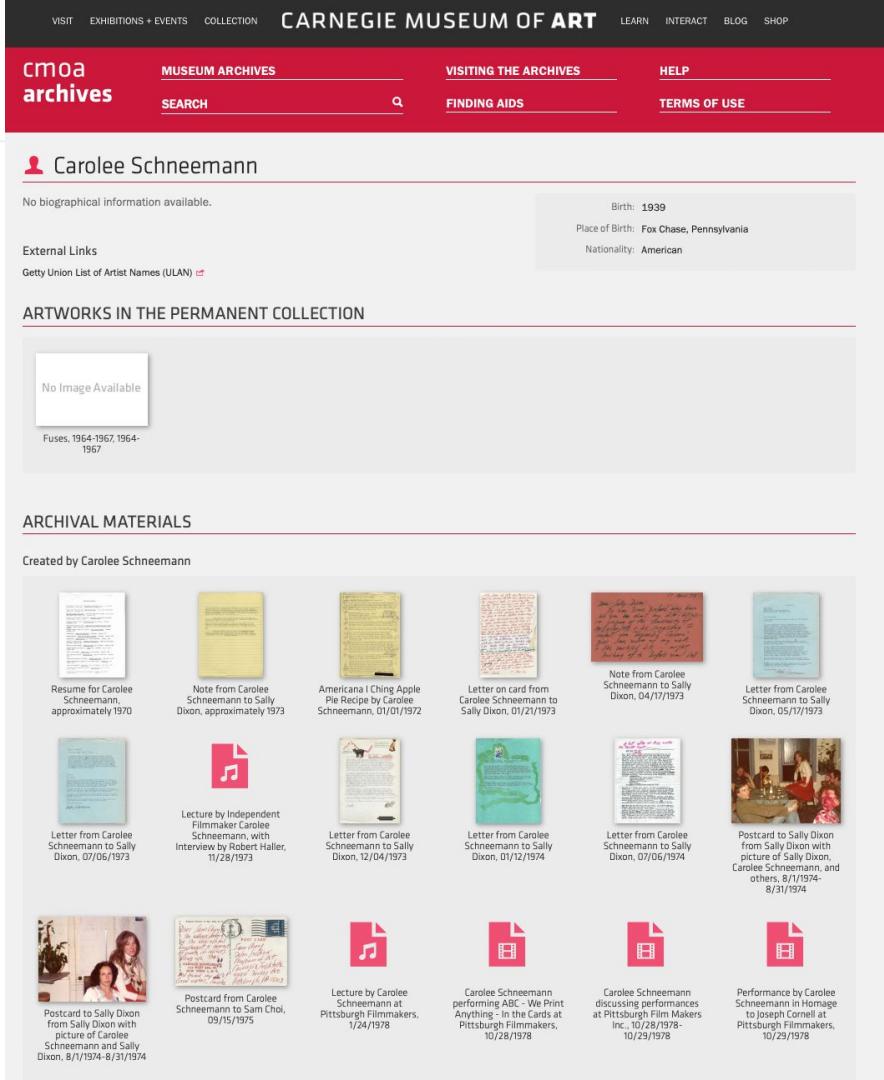
Getty

Laying the Groundwork: **Carnegie Museum Archives**

The archives of the Carnegie Museum of Art's Film Department launched in 2014.

The animating question was:

> What would happen if you treated the **relationships** between events, archival material, people, and artwork as the essential element, **not the objects**?

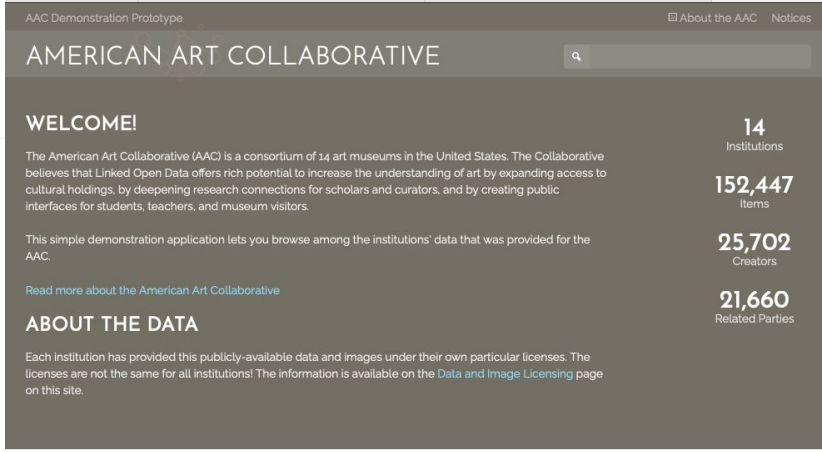


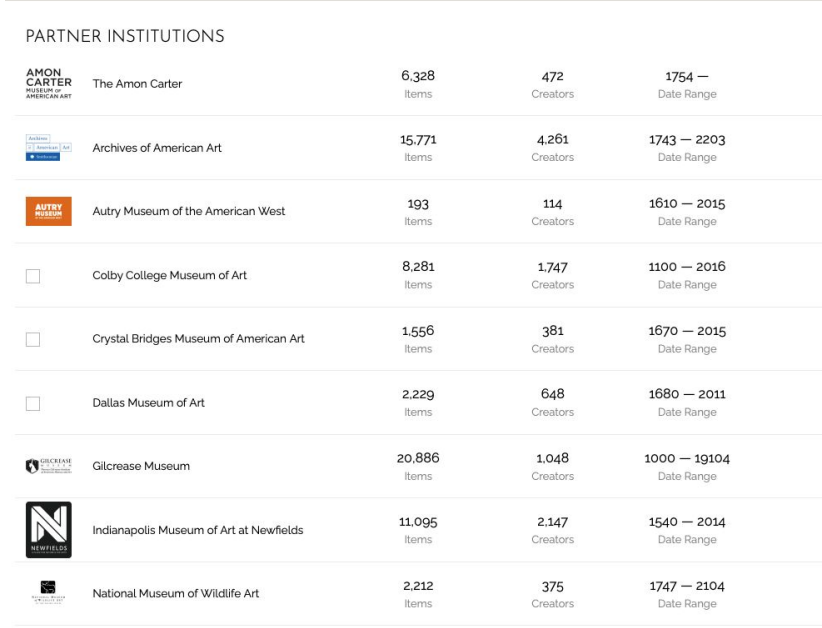
Getty

Laying the Groundwork: **American Art Collaborative**

In 2017, the American Art Collaborative launched.

It used these same principles to highlight connections across 14 institutions and 152,000 items—using Getty's Vocabularies as a bridging structure between institutions.



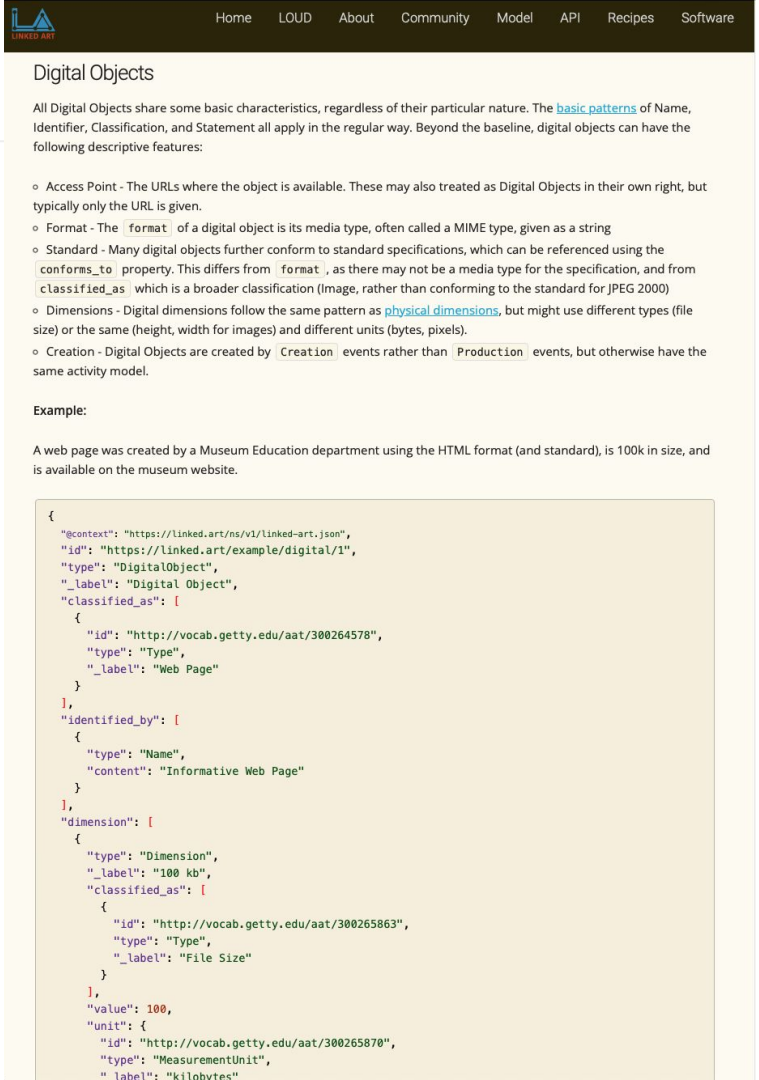


Getty

Laying the Groundwork: **Linked Art**

One of the most lasting outcomes of the American Art Collaborative was **Linked.Art**, the shared data model that connected institutions.

Getty



Home LOUD About Community Model API Recipes Software

Digital Objects

All Digital Objects share some basic characteristics, regardless of their particular nature. The basic patterns of Name, Identifier, Classification, and Statement all apply in the regular way. Beyond the baseline, digital objects can have the following descriptive features:

- Access Point - The URLs where the object is available. These may also treated as Digital Objects in their own right, but typically only the URL is given.
- Format - The `format` of a digital object is its media type, often called a MIME type, given as a string
- Standard - Many digital objects further conform to standard specifications, which can be referenced using the `conforms_to` property. This differs from `format`, as there may not be a media type for the specification, and from `classified_as` which is a broader classification (Image, rather than conforming to the standard for JPEG 2000)
- Dimensions - Digital dimensions follow the same pattern as physical dimensions, but might use different types (file size) or the same (height, width for images) and different units (bytes, pixels).
- Creation - Digital Objects are created by `Creation` events rather than `Production` events, but otherwise have the same activity model.

Example:

A web page was created by a Museum Education department using the HTML format (and standard), is 100k in size, and is available on the museum website.


```
{
  "@context": "https://linked.art/ns/v1/linked-art.json",
  "id": "https://linked.art/example/digital/1",
  "type": "DigitalObject",
  "_label": "Digital Object",
  "classified_as": [
    {
      "id": "http://vocab.getty.edu/aat/300264578",
      "type": "Type",
      "_label": "Web Page"
    }
  ],
  "identified_by": [
    {
      "type": "Name",
      "content": "Informative Web Page"
    }
  ],
  "dimension": [
    {
      "type": "Dimension",
      "_label": "100 kb",
      "classified_as": [
        {
          "id": "http://vocab.getty.edu/aat/300265863",
          "type": "Type",
          "_label": "File Size"
        }
      ],
      "value": 100,
      "unit": {
        "id": "http://vocab.getty.edu/aat/300265870",
        "type": "MeasurementUnit",
        "_label": "kilobytes"
```

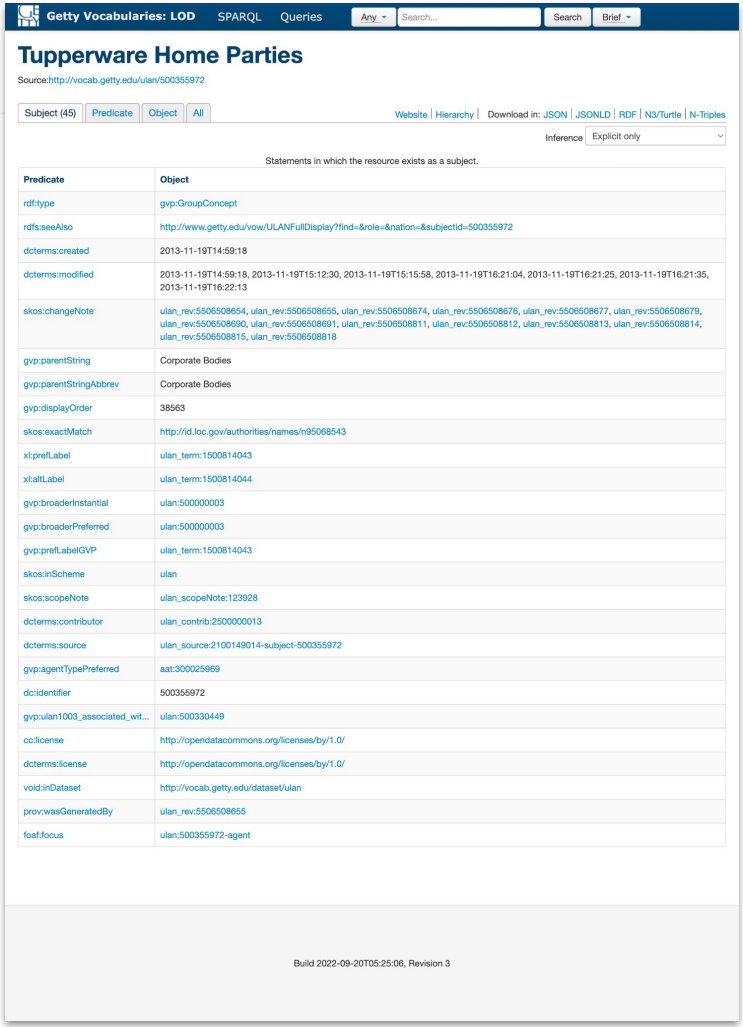
What we've done:

Getty's Digital Ecosystem

Getty

Getty has been doing Linked Data since 2014, starting with the Getty Vocabularies.

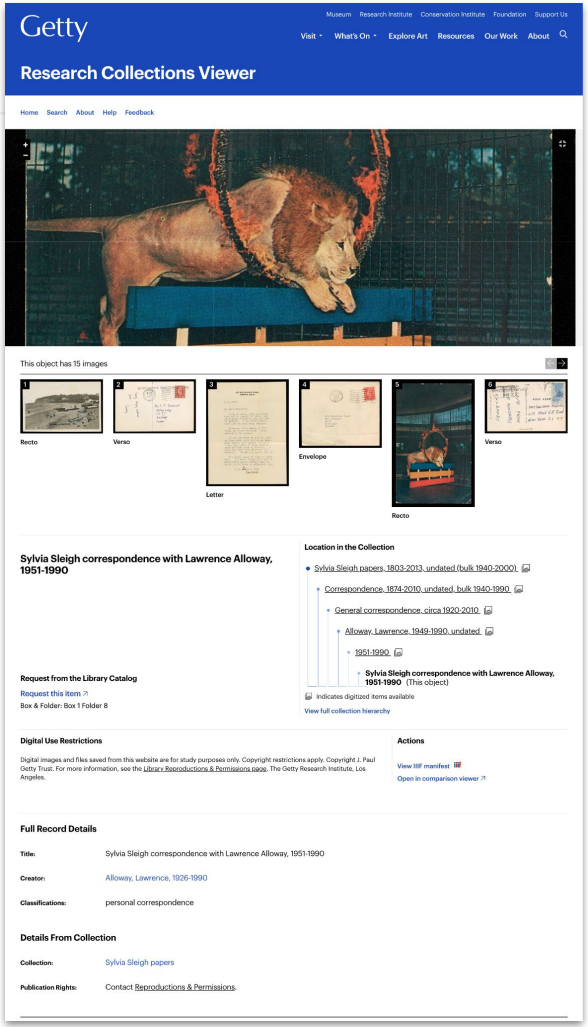
It's a thesaurus of concepts, people, and places used for cataloging across many institutions.



Getty

Getty's Linked Data: **Archival Records**

Since then, we've moved most of our major systems to use Linked Data—including our archives…

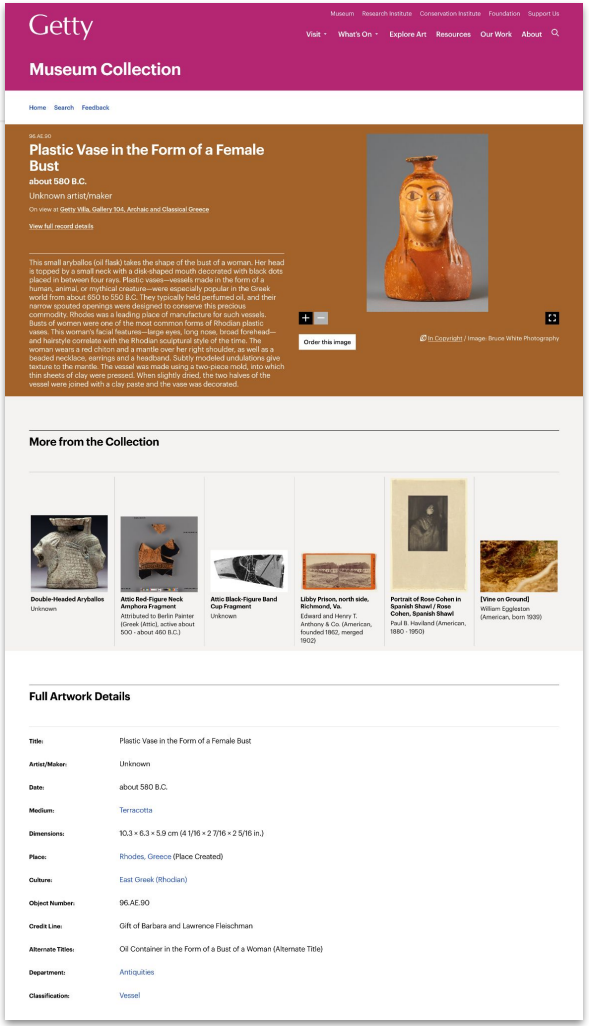


Getty

Getty's Linked Data: **Archival Records**

Since then, we've moved most of our major systems to use Linked Data—including our archives…

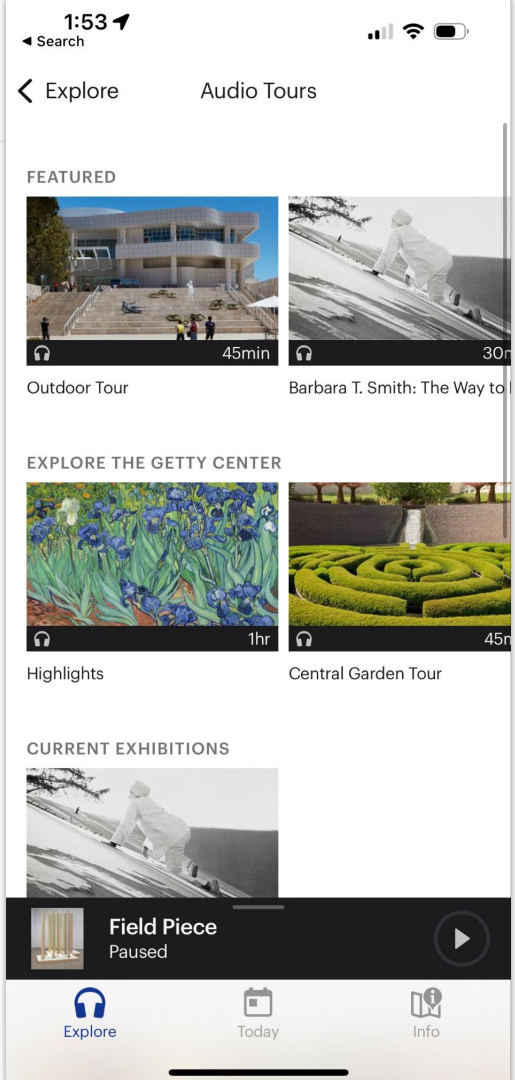
… and our museum collection.



Getty

Getty's Linked Data: **Audio Guide**

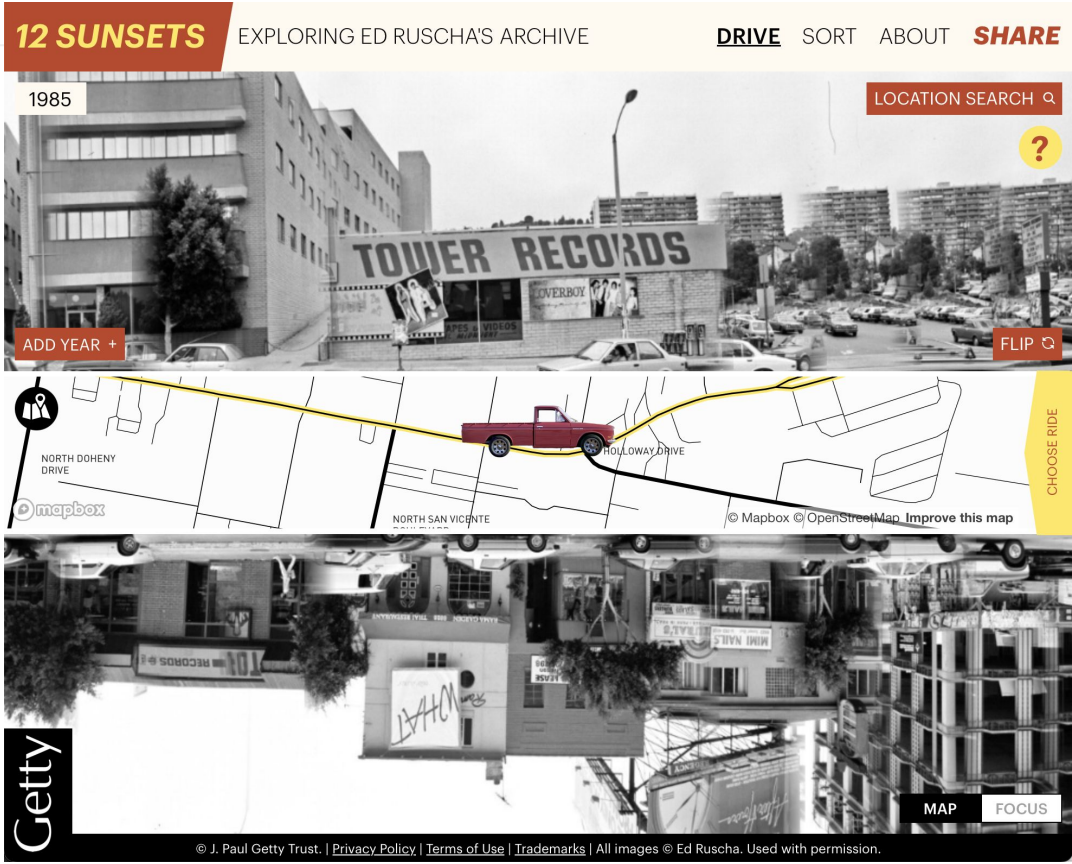
It's used for onsite visitor experiences via our audio guide…



Getty

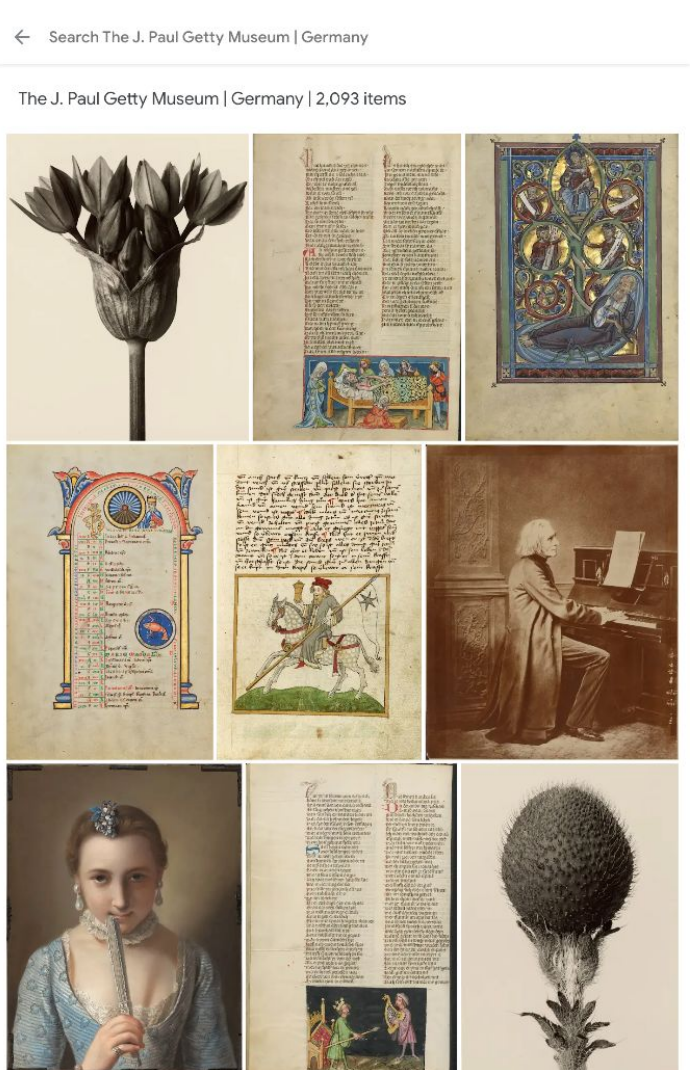
It's used for onsite visitor experiences
via our audio guide…

…and to provide novel interfaces for exploration of our materials.



Getty's Linked Data: **Google Arts & Culture**

It's also used by third parties: both large, like Google Arts & Culture…

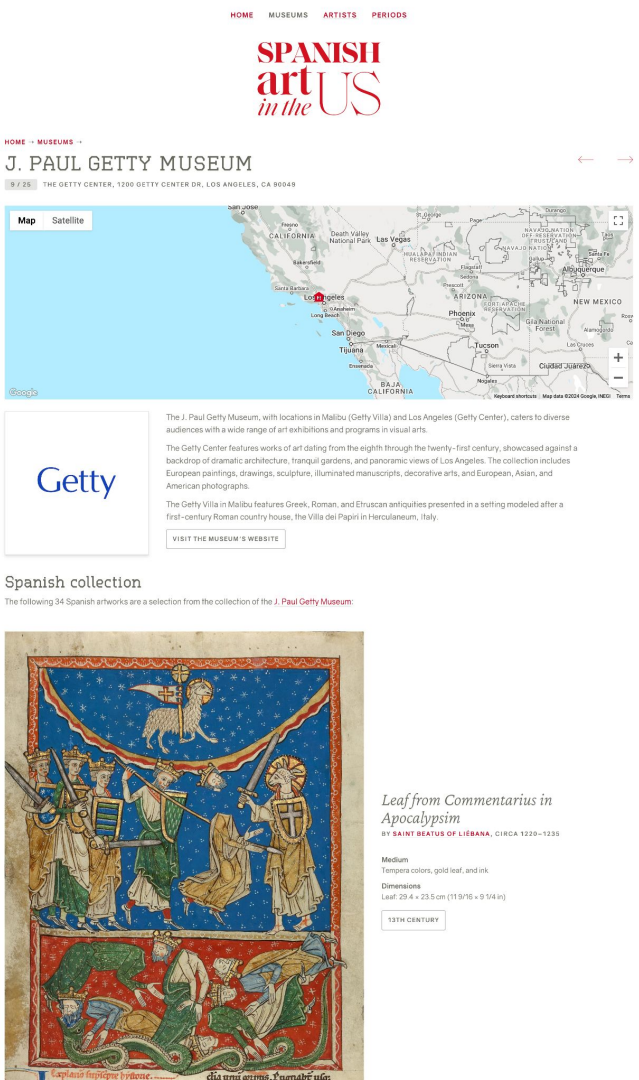


Getty

Getty's Linked Data: **Spanish Art in the US**

It's also used by third parties: both large, like Google Arts & Culture…

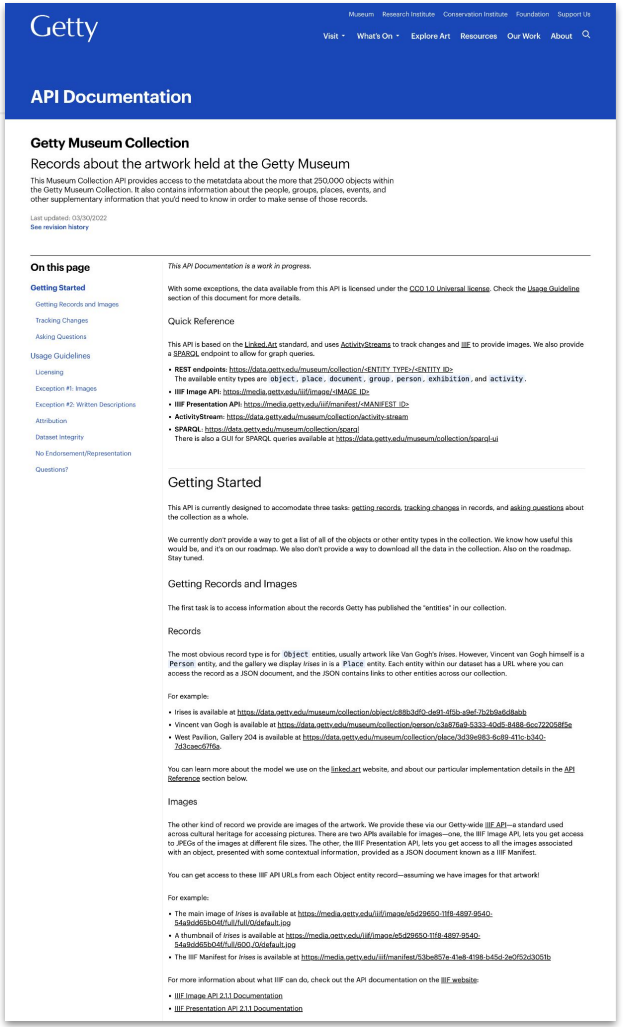
…and small, like this project by the Cultural Office of the Embassy of Spain.



Getty

Getty's Linked Data: **APIs**

We've also built a complex, powerful digital infrastructure to support this work—millions of records in a single shared data model, pulling from a wide collection of systems of record.



Getty

Museum Research Institute Conservation Institute Foundation Support Us

Visit What's On Explore Art Resources Our Work About

API Documentation

Getty Museum Collection

Records about the artwork held at the Getty Museum

This Museum Collection API provides access to the metatdata about the more that 250,000 objects within the Getty Museum Collection. It also contains information about the people, groups, places, events, and other supplementary information that you'd need to know in order to make sense of those records.

Last updated: 03/30/2022
See revision history

On this page

Getting Started
- Getting Records and Images
- Tracking Changes
- Asking Questions

Usage Guidelines
- Licensing
- Exception #1: Images
- Exception #2: Written Descriptions
- Attribution
- Dataset Integrity
- No Endorsement/Representation
- Questions?

This API Documentation is a work in progress.

With some exceptions, the data available from this API is licensed under the CCO 1.0 Universal license. Check the Usage Guideline section of this document for more details.

Quick Reference

This API is based on the Linked.Art standard, and uses ActivityStreams to track changes and IIIF to provide images. We also provide a SPARQL endpoint to allow for graph queries.

- **REST endpoints**: https://data.getty.edu/museum/collection/<ENTITY TYPE>/<ENTITY ID>
  The available entity types are `object`, `place`, `document`, `group`, `person`, `exhibition`, and `activity`.
- **IIIF Image API**: https://media.getty.edu/iiif/image/<IMAGE ID>
- **IIIF Presentation API**: https://media.getty.edu/iiif/manifest/<MANIFEST ID>
- **ActivityStream**: https://data.getty.edu/museum/collection/activity-stream
- **SPARQL**: https://data.getty.edu/museum/collection/sparql
  There is also a GUI for SPARQL queries available at https://data.getty.edu/museum/collection/sparql-ui

Getting Started

This API is currently designed to accomodate three tasks: getting records, tracking changes in records, and asking questions about the collection as a whole.

We currently don't provide a way to get a list of all of the objects or other entity types in the collection. We know how useful this would be, and it's on our roadmap. We also don't provide a way to download all the data in the collection. Also on the roadmap. Stay tuned.

Getting Records and Images

The first task is to access information about the records Getty has published the "entities" in our collection.

Records

The most obvious record type is for `Object` entities, usually artwork like Van Gogh's *Irises*. However, Vincent van Gogh himself is a `Person` entity, and the gallery we display *Irises* in is a `Place` entity. Each entity within our dataset has a URL where you can access the record as a JSON document, and the JSON contains links to other entities across our collection.

For example:

- Irises is available at https://data.getty.edu/museum/collection/object/c88b3df0-de91-4f5b-a9ef-7b2fd9a6d8abb
- Vincent van Gogh is available at https://data.getty.edu/museum/collection/person/e3a857ea9-5333-40d5-8488-6cc72205815e
- West Pavilion, Gallery 204 is available at https://data.getty.edu/museum/collection/place/3d39e983-6c89-411c-b340-7d3caec67f6a.

You can learn more about the model we use on the linked.art website, and about our particular implementation details in the API Reference section below.

Images

The other kind of record we provide are images of the artwork. We provide these via our Getty-wide IIIF API—a standard used across cultural heritage for accessing pictures. There are two APIs available for images—one, the IIIF Image API, lets you get access to JPEGs of the images at different file sizes. The other, the IIIF Presentation API, lets you get access to all the images associated with an object, presented with some contextual information, provided as a JSON document known as a IIIF Manifest.

You can get access to these IIIF API URLs from each Object entity record—assuming we have images for that artwork!

For example:

- The main image of *Irises* is available at https://media.getty.edu/iiif/image/e5d29650-11f8-4897-9540-54a9dd65b04f/full/full/0/default.jpg
- A thumbnail of *Irises* is available at https://media.getty.edu/iiif/image/e5d29650-11f8-4897-9540-54a9dd65b04f/full/500,0/default.jpg
- The IIIF Manifest for *Irises* is available at https://media.getty.edu/iiif/manifest/53be857e-41e8-4198-b45d-2e0f52d3051b

For more information about what IIIF can do, check out the API documentation on the IIIF website:

- IIIF Image API 2.1.1 Documentation
- IIIF Presentation API 2.1.1 Documentation

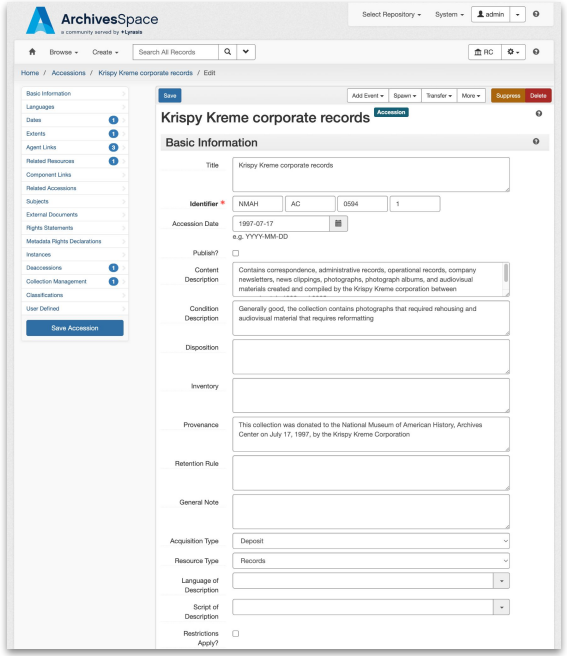
Getty

Under the hood:

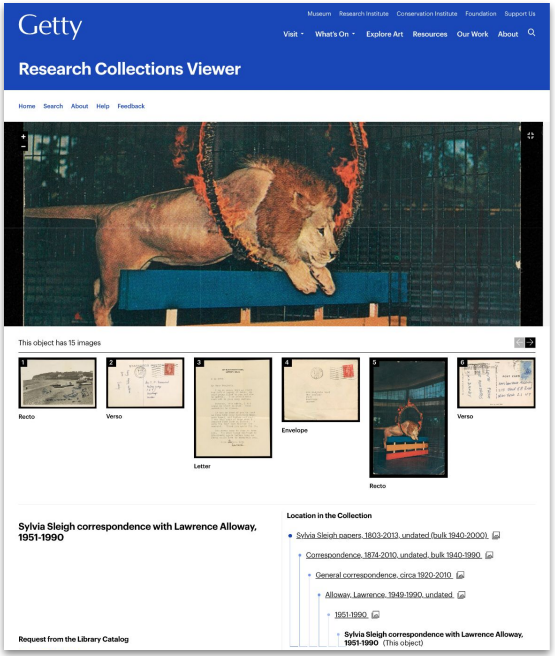
Linked Data & the Everything API

Getty

Data Flow: **How we wish it was**

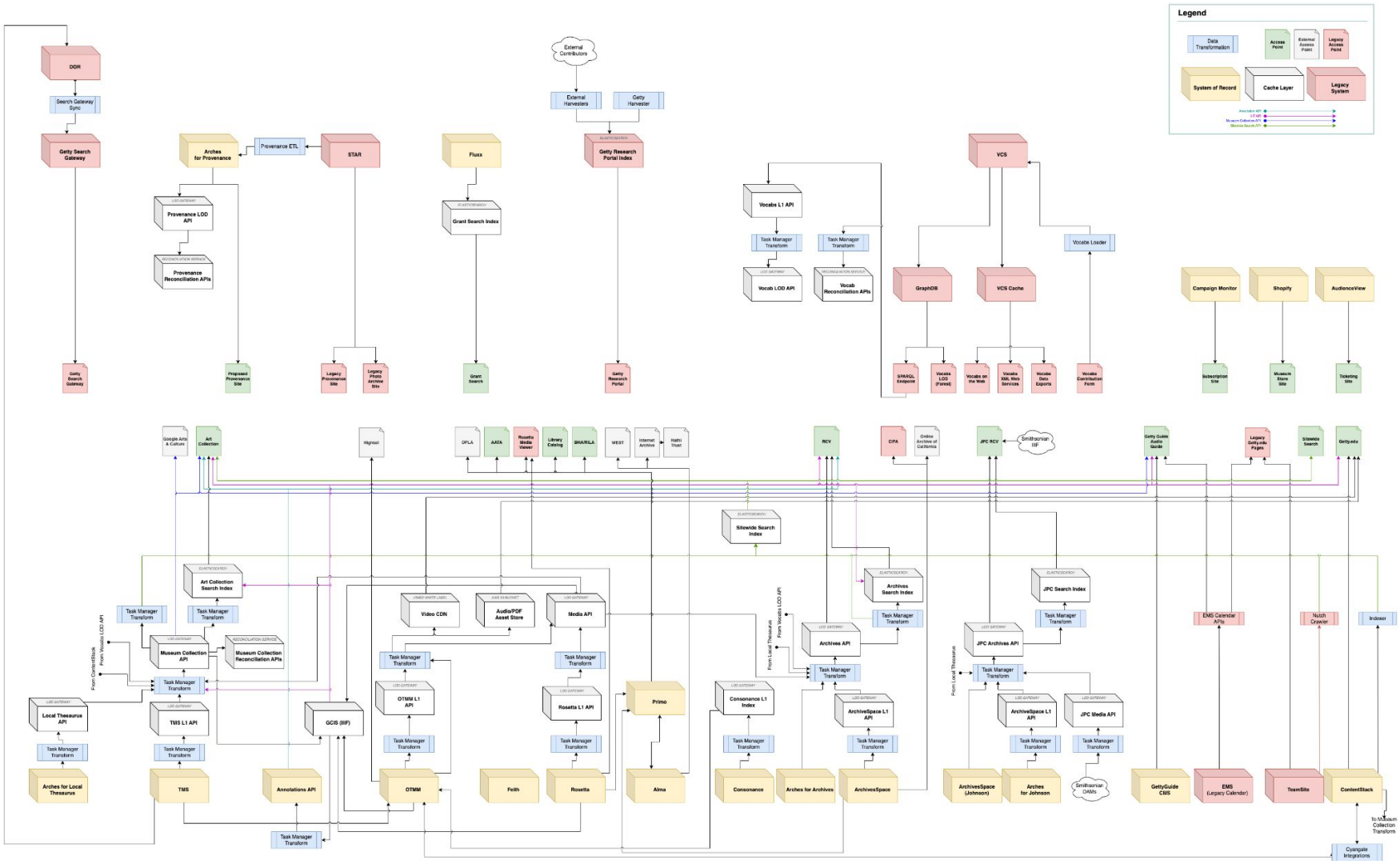


Staff Interface



Public Website

Getty

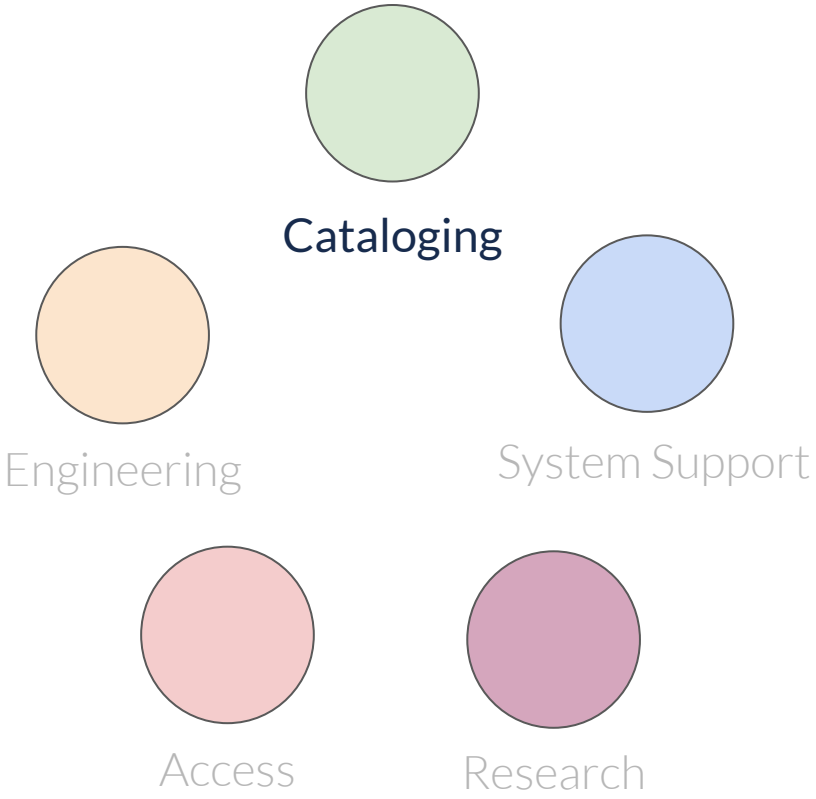

Legend
Data Transformation
Access Point
External Access Point
Legacy Access Point
System of Record
Cache Layer
Legacy System
Annotation API
IIIF API
Museum Collection API
Glossia Search API
DDR
Search Gateway Sync
Getty Search Gateway
Getty Search Gateway
Arches for Provenance
Provenance ETL
STAR
Provenance LOD API
Provenance Reconciliation APIs
Proposed Provenance Site
Legacy Provenance Site
Legacy Photo Archive Site
Fluxx
Grant Search Index
Grant Search
External Contributors
External Harvesters
Getty Harvester
Getty Research Portal Index
Getty Research Portal
Vocabs L1 API
Task Manager Transform
Task Manager Transform
Vocab LOD API
Vocab Reconciliation APIs
VCS
Vocabs Loader
GraphDB
VCS Cache
SPARQL Endpoint
Vocabs LOD (Forest)
Vocabs on the Web
Vocabs XML Web Services
Vocabs Data Exports
Vocabs Contribution Form
Campaign Monitor
Shopify
AudienceView
Subscription Site
Museum Store Site
Ticketing Site
Google Arts & Culture
Art Collection
Hipmail
DPLA
AATA
Rosetta Media Viewer
Library Catalog
BHA/RLA
WEST
Internet Archive
Hathi Trust
RCV
CIPA
Online Archive of California
JPC RCV
Smithsonian IIIF
Getty Guide Audio Guide
Legacy Getty.edu Pages
Sitewide Search
Getty.edu
Sitewide Search Index
Art Collection Search Index
Task Manager Transform
Task Manager Transform
From Content/Risk
From Vocabs LOD API
Museum Collection API
Museum Collection Reconciliation APIs
Task Manager Transform
Local Thesaurus API
Task Manager Transform
Arches for Local Thesaurus
TMS L1 API
Task Manager Transform
TMS
GCIS (IIIF)
Task Manager Transform
Annotations API
Video CDN
Audio/PDF Asset Store
Media API
Task Manager Transform
OTMM L1 API
Task Manager Transform
OTMM
Task Manager Transform
Rosetta L1 API
Task Manager Transform
Feith
Rosetta
Primo
Alma
Archives Search Index
Task Manager Transform
From Local Thesaurus
From Vocabs LOD API
Archives API
Task Manager Transform
Consonance L1 Index
Task Manager Transform
Consonance
ArchiveSpace L1 API
Task Manager Transform
Arches for Archives
ArchivesSpace
JPC Search Index
Task Manager Transform
From Local Thesaurus
JPC Archives API
Task Manager Transform
ArchiveSpace L1 API
Task Manager Transform
ArchivesSpace (Johnson)
JPC Media API
Task Manager Transform
Arches for Johnson
Smithsonian DAMs
GettyGuide CMS
EMS Calendar APIs
EMS (Legacy Calendar)
Notch Crawler
TeamSite
Indexer
ContentStack
To Museum Collection Transform
Cyangate Integrations

These systems support people.

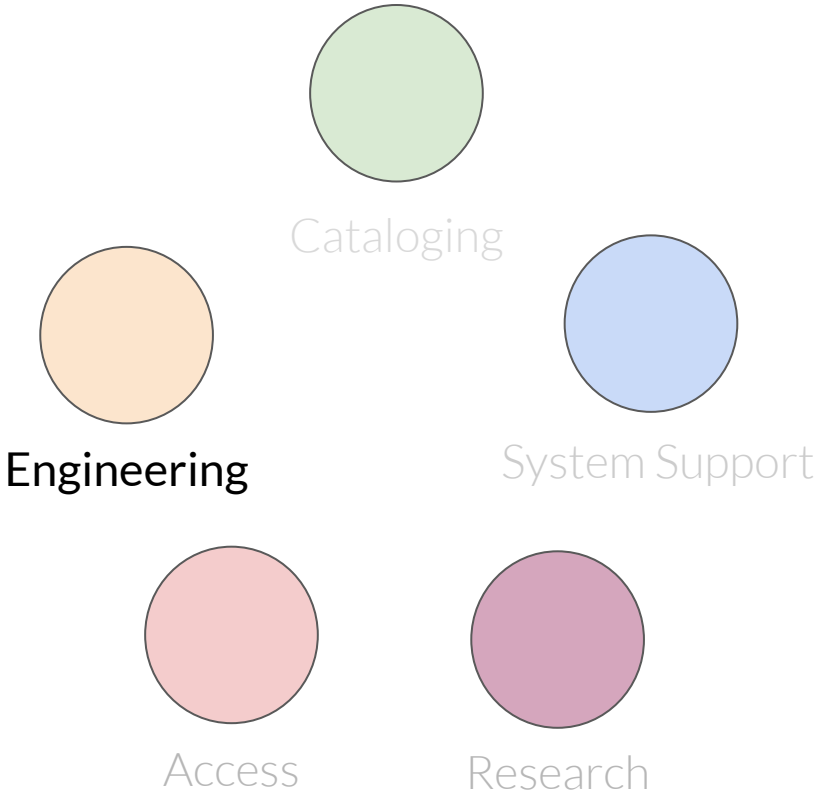
Digital infrastructure is designed to use computers to empower **people** to be more effective at meeting the mission of the organization.

Catalogers need systems that match their workflows—and different disciplines have different needs.

Our infrastructure needed to not be tied to any particular backend system.

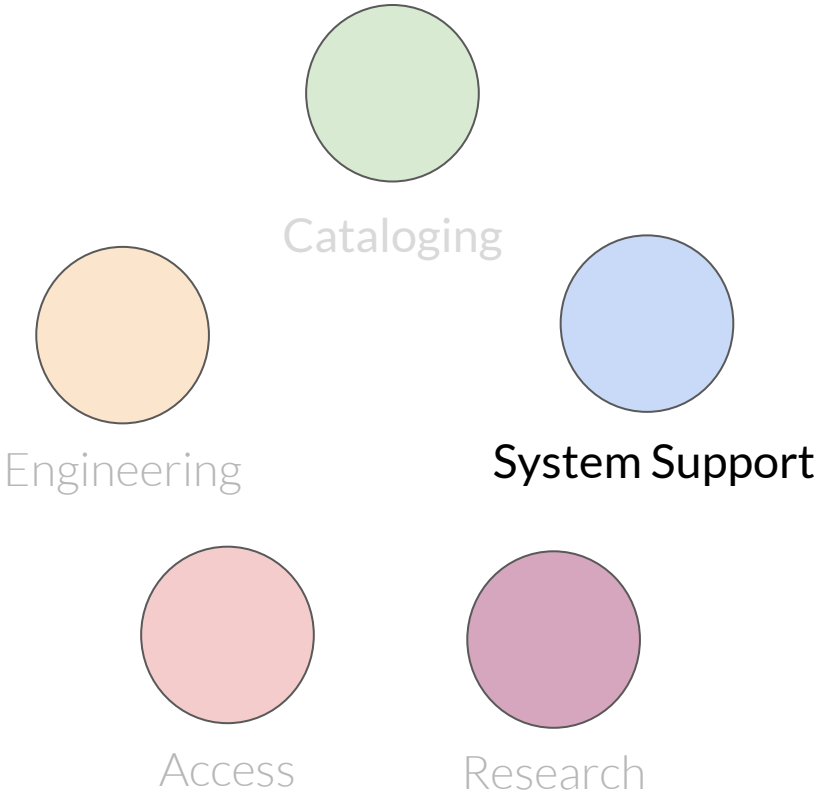


Engineers need to get data in and out of systems, using patterns and practices that they already know how to use.



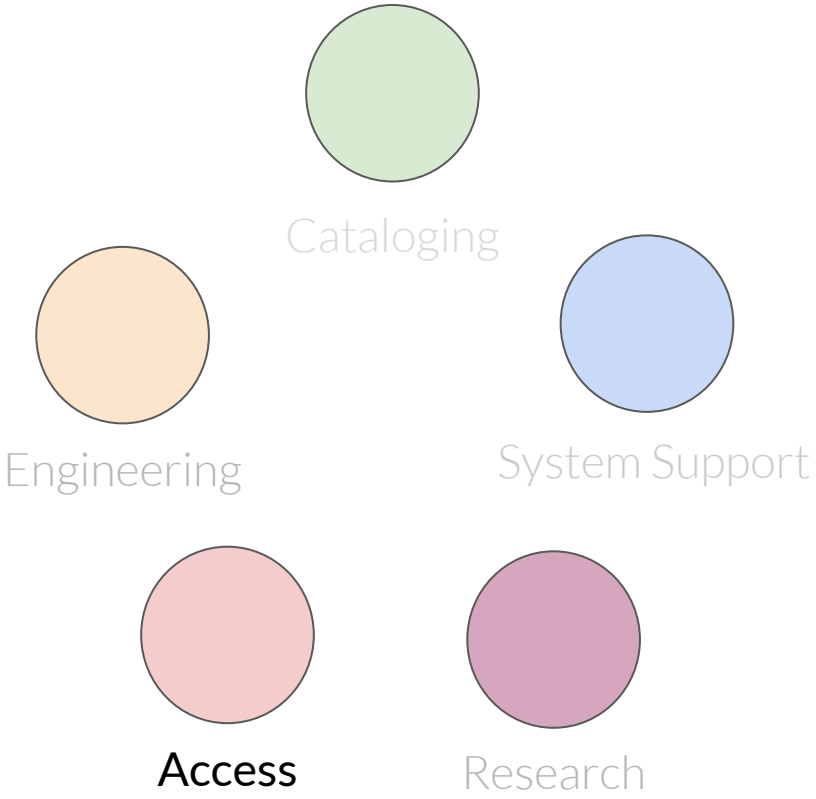
System admins just don't want you to break their stuff.

Pulling data out of systems on demand usually breaks stuff.

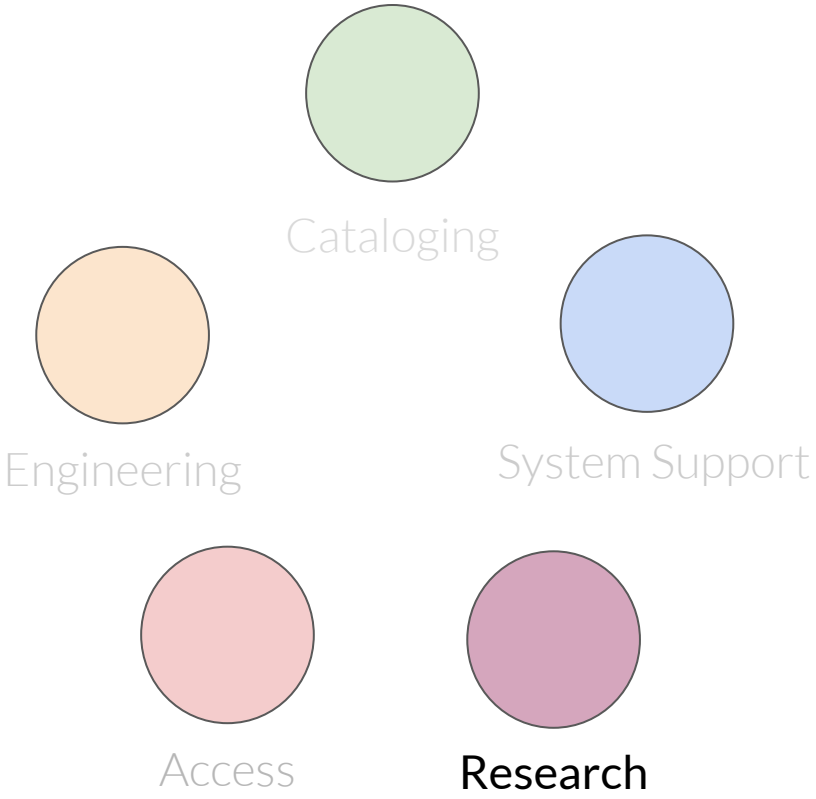


Most end users are looking for content—they want to learn what we know on a given topic.

This may be professional scholarship or it might be looking for pictures—both are examples of information-seeking behaviours.



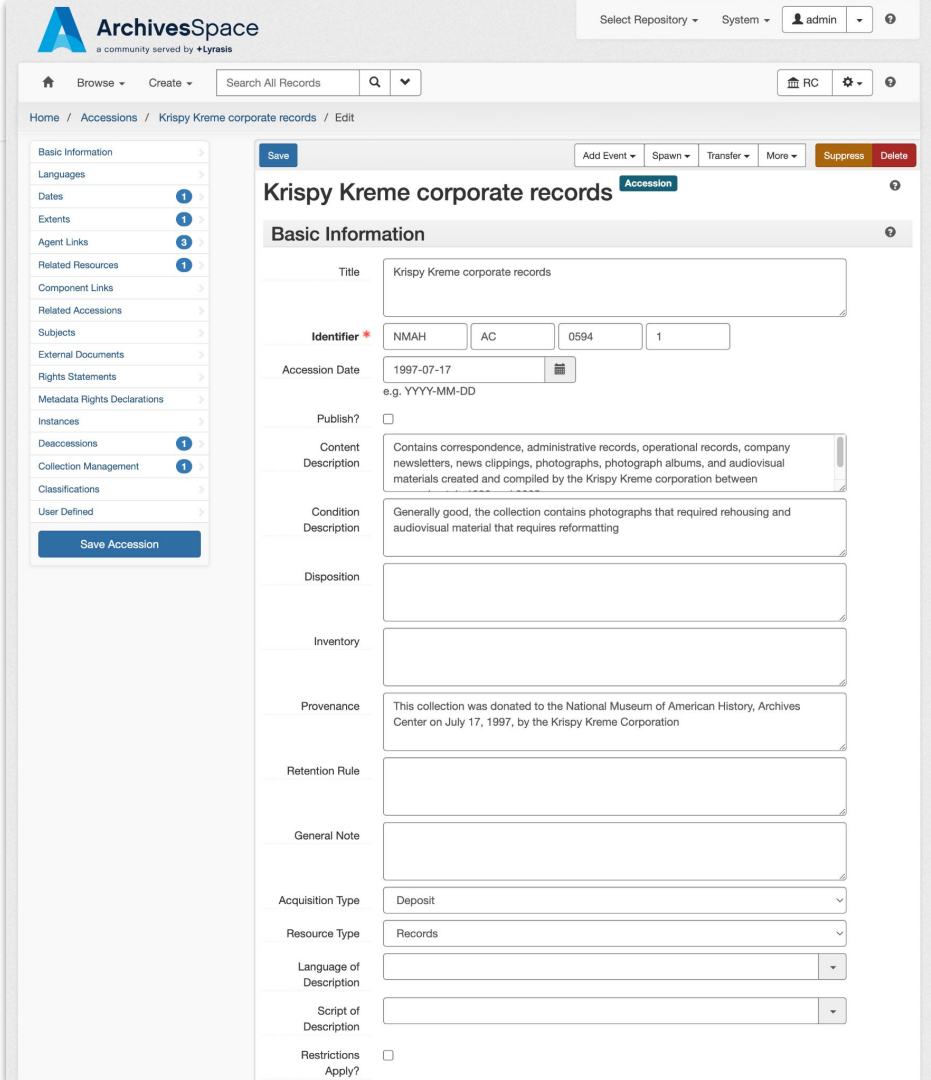
And some researchers want to find questions that haven't been asked before—to find new connections or patterns in the data that others have overlooked.



Getty's Linked Data: **The LOD Gateway**

Meeting the needs of catalogers is *mostly* not my problem.

There are high-quality, professional tools that work within the disciplinary training of the field.



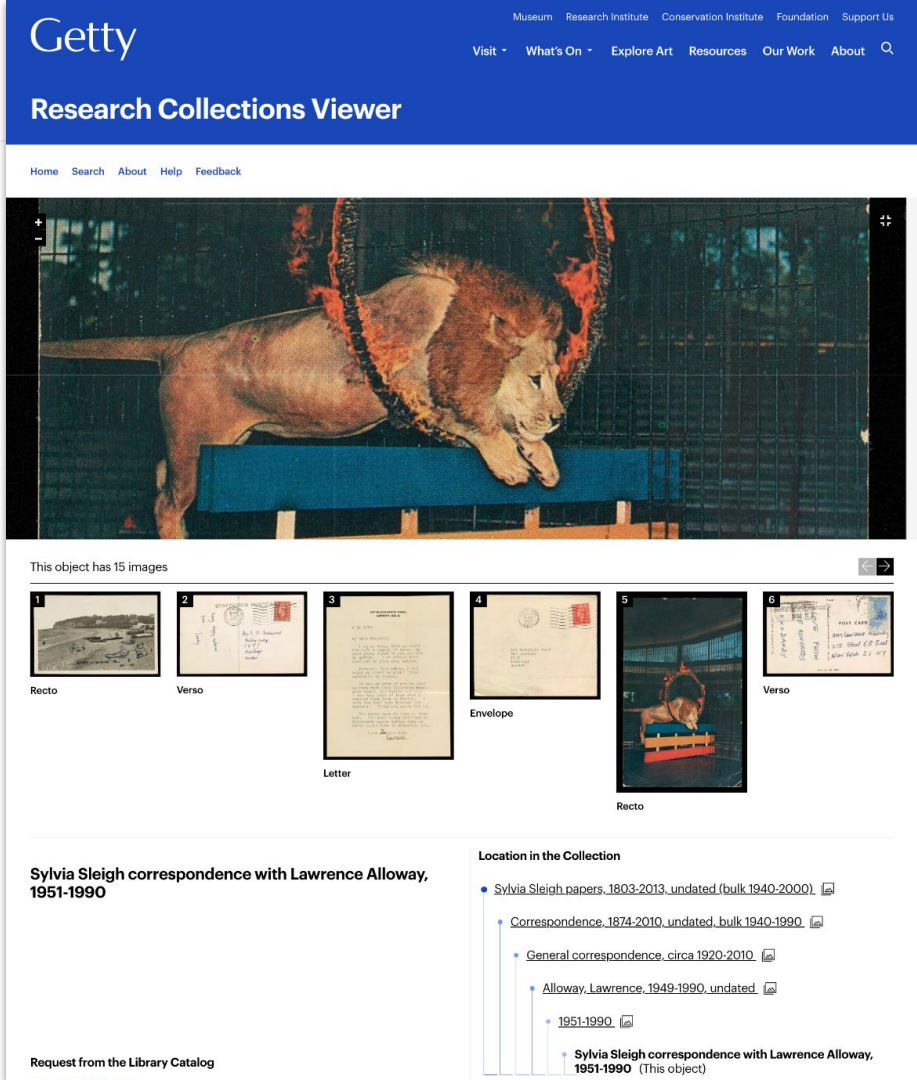
Getty

Getty's Linked Data: **The LOD Gateway**

Providing access to that data, though, often requires recontextualization:

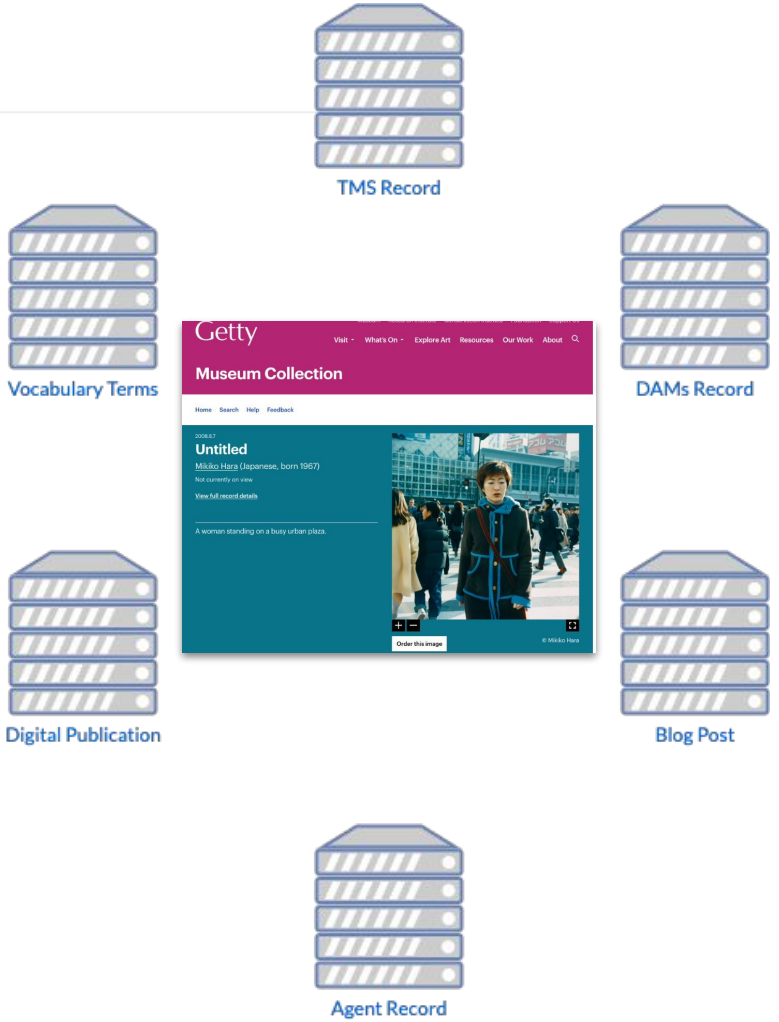
Changing the conceptual lens from one focused on staff efficiencies to one focused on user's needs.

Getty

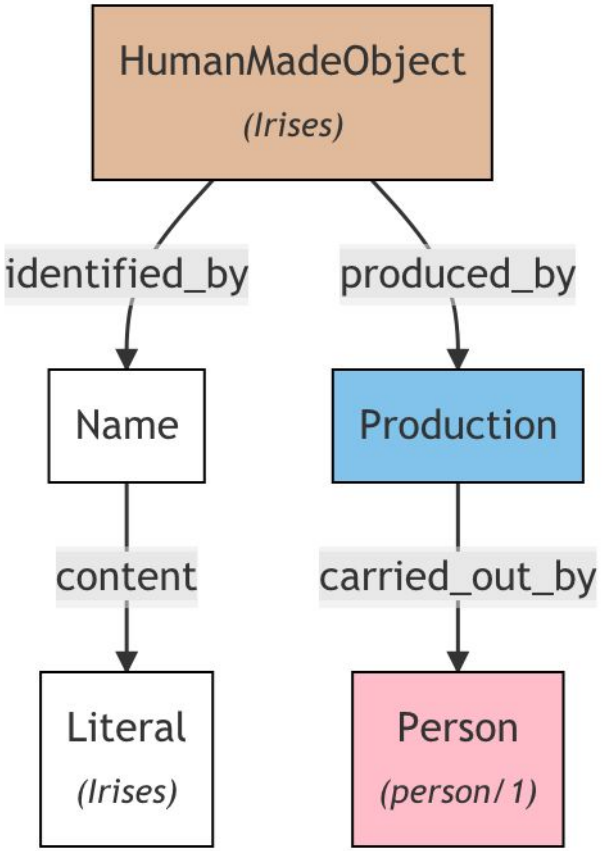


Doing so requires combining data from multiple systems and multiple workflows into a new record.

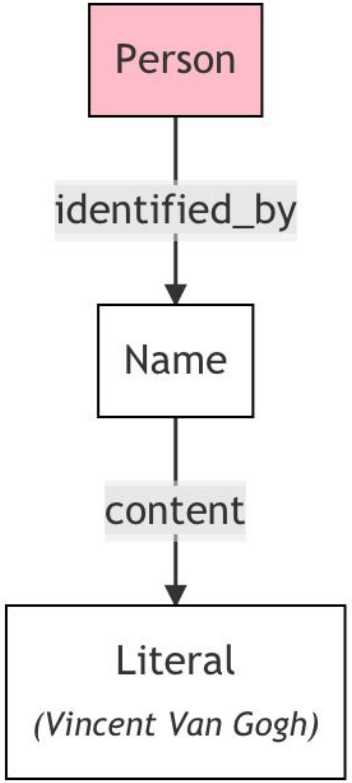
This combining—or linking—of data has tradeoffs.



Imagine a record for the painting ***Irises***.



And a second record, this one for **Van Gogh**.

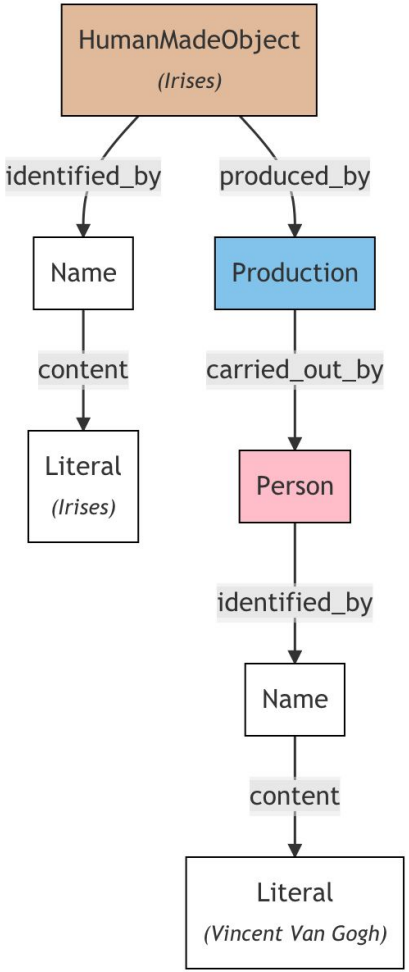


These could be seen as two separate documents:

```
"@context":
"https://linked.art/ns/v1/linked-art.json",
"id": "person/1",
"type": "Person",
"identified_by": {
      "id": "person/1/name",
      "type": "Name",
      "content": "Vincent Van Gogh"
}
```

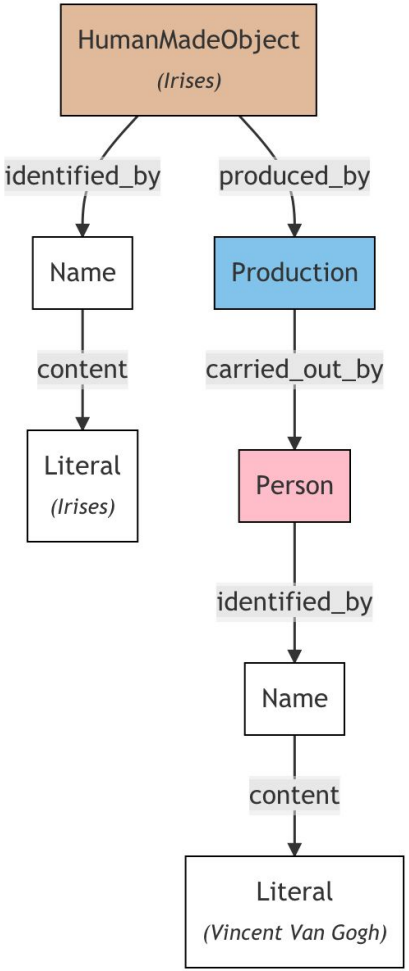
```
"@context":
"https://linked.art/ns/v1/linked-art.json",
"id": "object/1",
"type": "HumanMadeObject",
"identified_by": {
   "id": "object/1/name",
   "type": "Name",
   "content": "Irises"
},
"produced_by": {
   "id": "object/1/production",
   "carried_out_by": {"id":"person/1"}
}
```

Or as a single graph.



From the point of view of the data, these are equivalent—they contain the same facts.

But from a usability perspective, they make different things easy or hard.



Getty

Documents are optimized for **Access**:

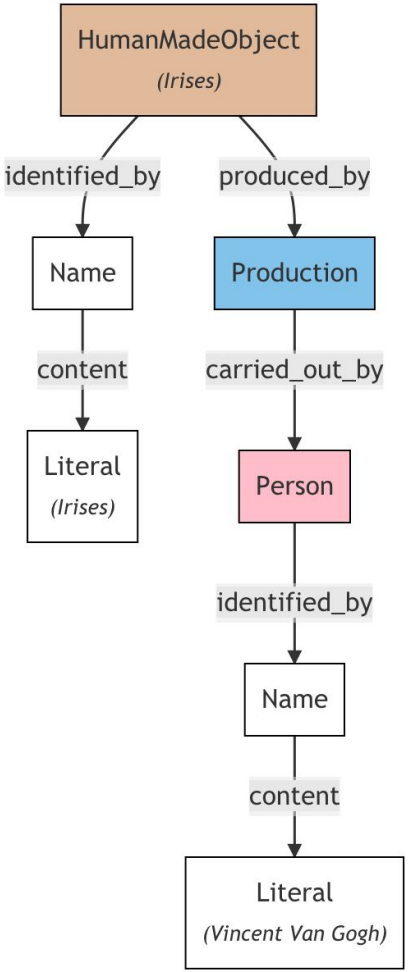
They provide a specific set of data bundled together by the data creator that provide all the facts you need…given a **specific context**.

```
"@context":
"https://linked.art/ns/v1/linked-art.json",
"id": "object/1",
"type": "HumanMadeObject",
"identified_by": {
   "id": "object/1/name",
   "type": "Name",
   "content": "Irises"
},
"produced_by": {
   "id": "object/1/production",
   "carried_out_by": {"id":"person/1"}
}
```

Getty

Graphs, alternately, are optimized for **querying**:

Allowing a user to define a specific context based on novel criteria and returning that subset of facts.



Imagine two Questions:

“What objects does Getty have that have images larger than 1200px on the longest side that have been exhibited in both New York and Paris and were created by artists who lived before 1850?

and

What’s the label info for *Irises*?

Getty

Imagine two Questions:

At the Getty, we have *never* asked:

“What objects does Getty have that have images larger than 1200px on the longest side that have been exhibited in both New York and Paris and were created by artists who lived before 1850?

but we ask

What’s the label info for *Irises*?

Several thousand times a day.

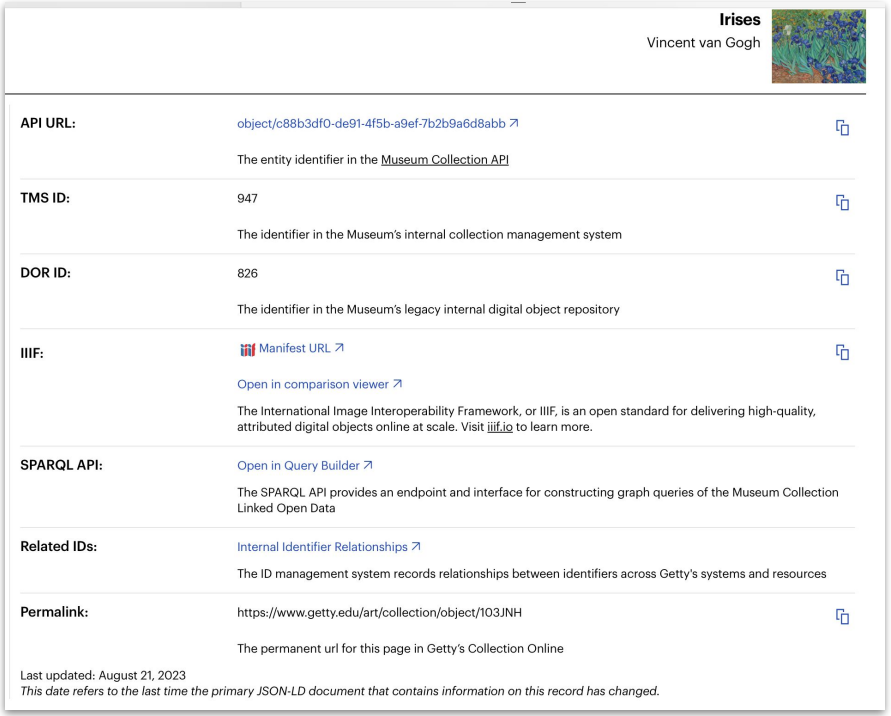
Having an interface for documents lets us provide a simple, easily understandable record that maps well to known contexts.

This is important, because people usually *expect* these contexts. It makes answering common questions simple.

Documents: **For Access and Discovery**

Documents are also the way **the internet** works: REST APIs, cache control, JSON, webpages.

Using these well-known systems helps **developers** make systems that are fast and easy to build.



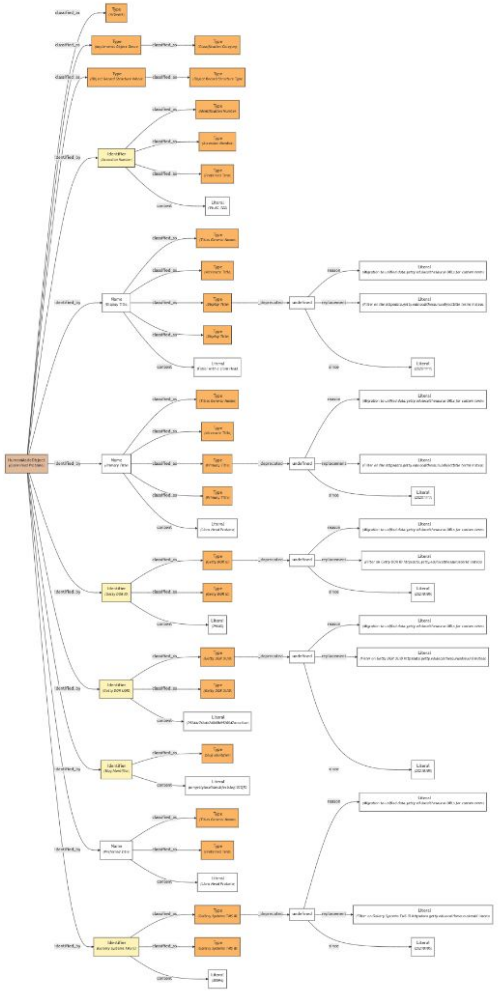
Getty

Research is different—each scholar brings their own question and their own context.

Meeting their need means empowering them to draw **their own** boundaries within the data.

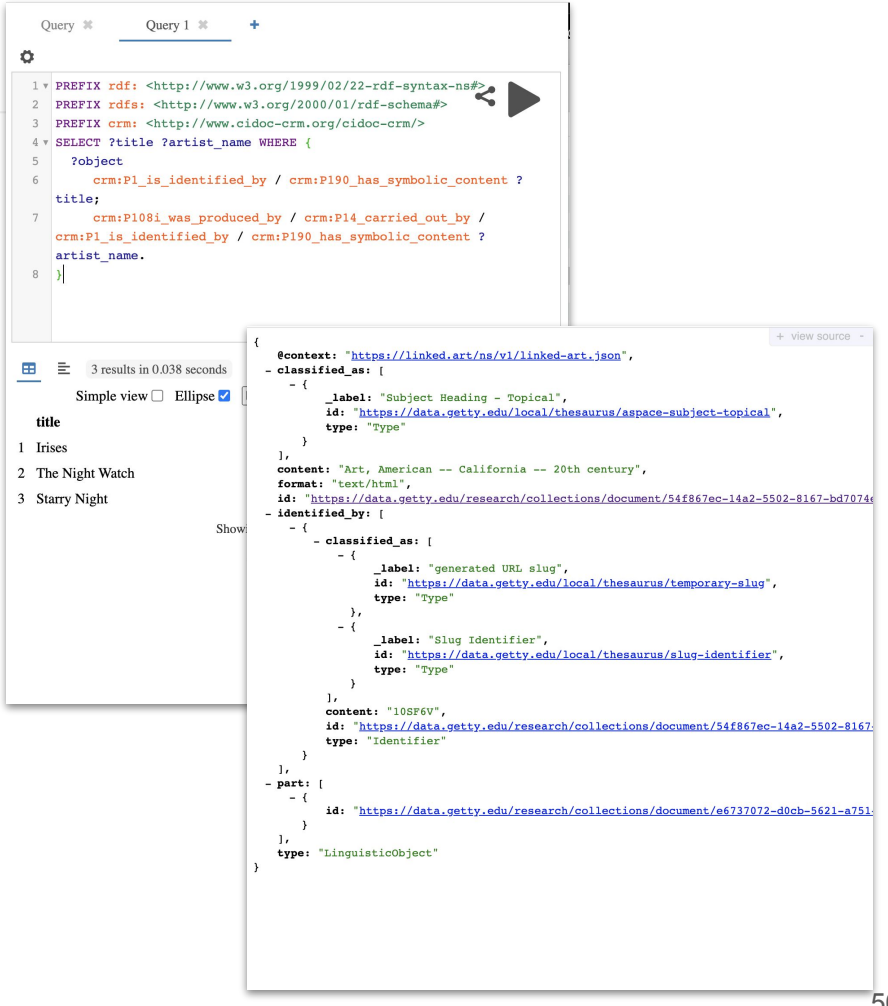
Doing so is complex—it moves the burden of defining the relevant context to the **user** of the data, not the **creator** of the data.

But it makes asking new questions **possible**, even if it might be inefficient or complicated.



We've built our infrastructure to allow for both use cases:

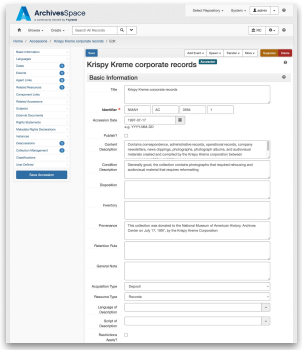
A developer can create, update, and delete documents, and behind the scenes it will keep a graph in sync with those changes.

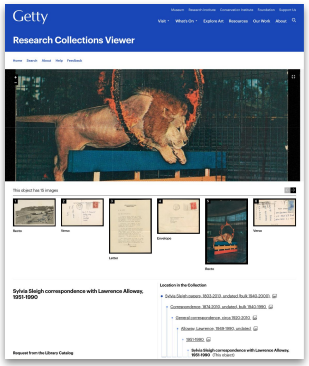


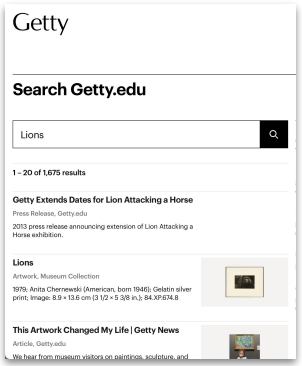
Getty

It also allows for synchronization across systems:

A editor changes a record, which means the API needs updated, which means the website needs updated, and the search interfaces, and third-party systems…

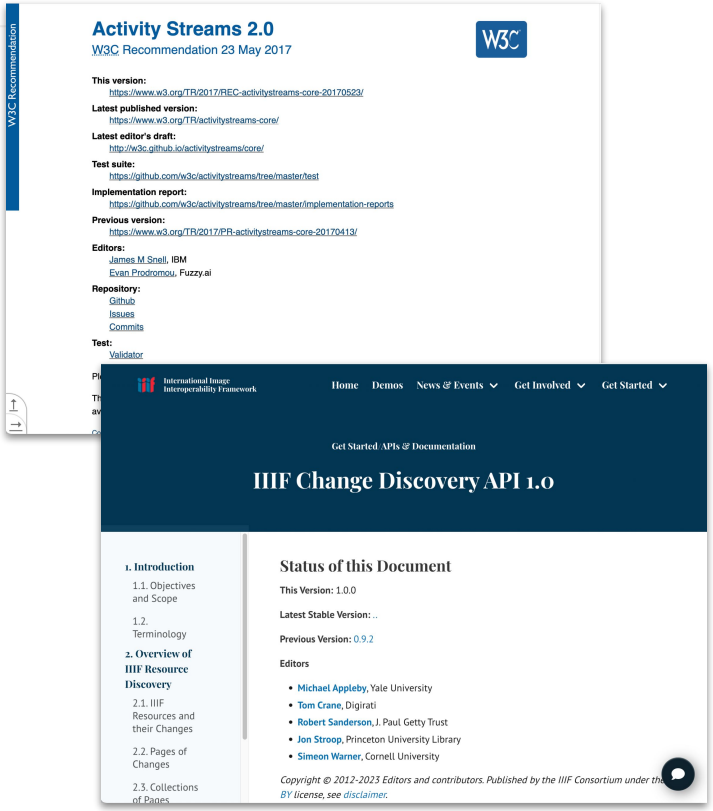






The infrastructure uses uses the **W3C ActivityStream** standard and are implemented using the patterns from the **IIIF Change Discovery API**.

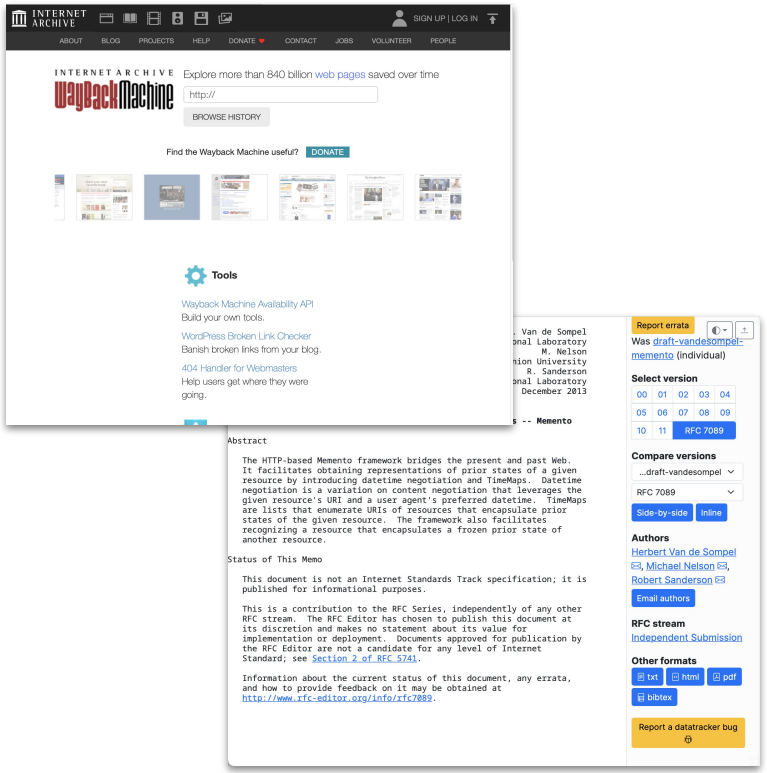
Using standards makes it easy to build integrations against changing data—both within our organization and for external aggregators.



Getty

For some kinds of data, it's also valuable to also be able to see **what has changed over time** for a given record.

To do so, our APIs also supports **Memento**, the standard underneath the Internet Archive.



Getty

This lets you automatically open older versions of the record—providing an **audit log** and the ability for scholars to understand how **knowledge changes over time**.

Getty

Cool Tech, Bro:

Why Does this Matter?

Getty

A Hard-won lesson:

No application that we've built **required** Linked Data.

A Hard-won lesson:

No application that we've built **required** Linked Data.

Which, if you think about it, makes sense. Each application has a **specific, known context** with clear record boundaries.

A Hard-won lesson:

Different users have different contexts and need different affordances.

A shared, graph-based data model **allows us to re-present the data** in a way that matches user's varying models of the world via multiple interfaces.

As Simple as possible:

A shared data model also makes our developers more effective—eventually.

Building on top of web technology lets the engineering learning curve be gradual.

As Simple as possible:

Standards are valuable for interoperability—but also because you don't have to **write all the documentation.**

Nobody wants to write it, but you can't work across institutions without it.

As Simple as possible:

Minimize complexity in the data model.

Data is for computers—text is for humans. Resist the urge to show off.

You can always add complexity—you can **never take it away**.

Disciplinary Misdeeds:

The hardest part of this will be **change management**.

Recontextualizing information across boundaries hides disciplinary labor—and digital innovations can conflict with pre-digital best practices.

Evangelize and collaborate.

What makes cultural data interesting is *not* contained within any one institution.

It's shared across our entire, world-wide community. We should work together.

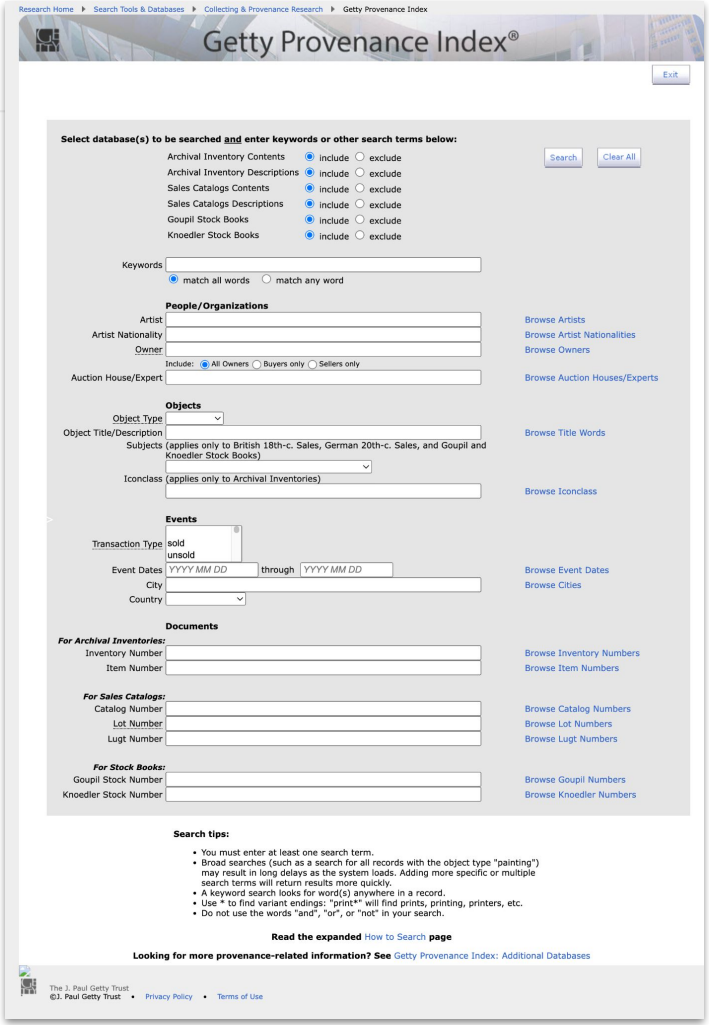
Shared models and shared code make that easier.

What's Next: **Provenance Index**

Over the next several years, we'll be expanding our usage of this system:

This fall, we'll launch a new version of the Getty **Provenance Index**, adding in 22M records of transactions between art dealers.

This **research-focused dataset** will allow new insights into collections around the world—and into the art market as a whole.

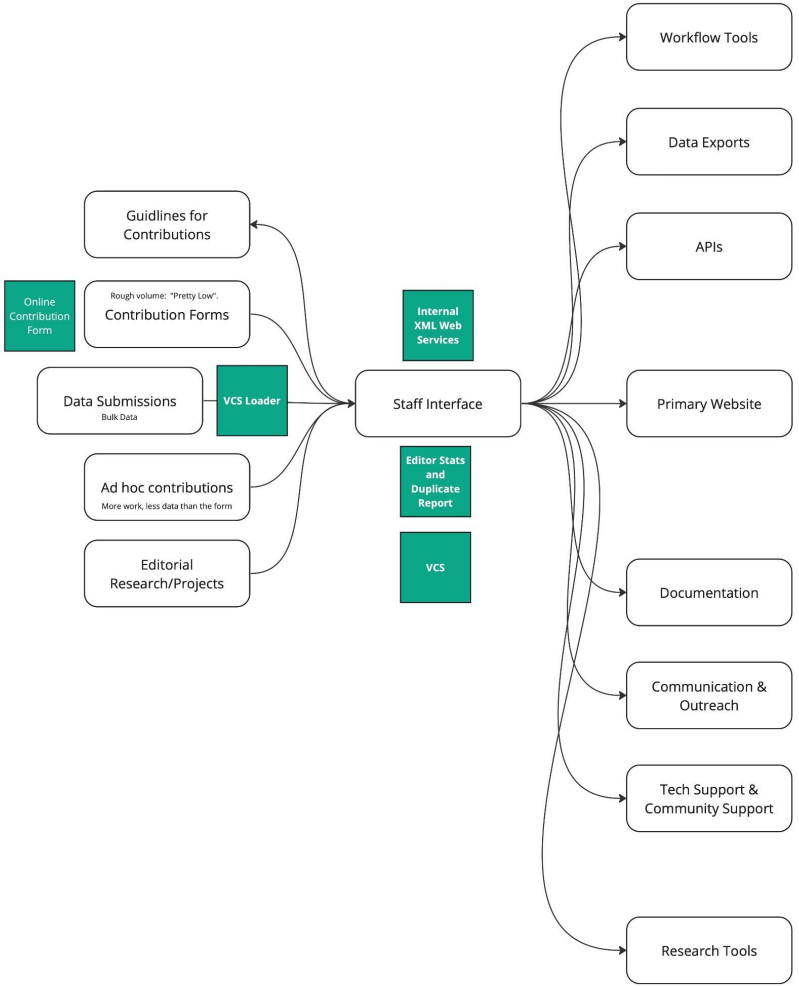


Getty

What's Next: **Getty Vocabularies**

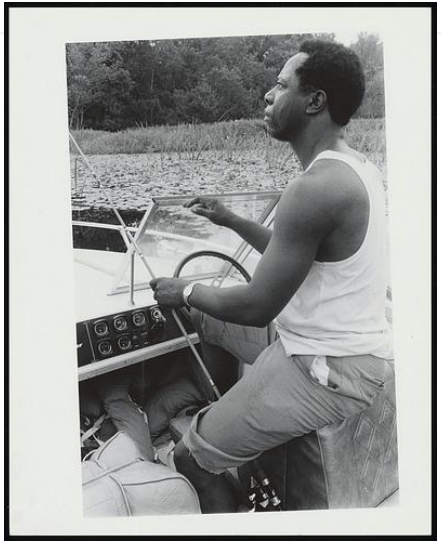
We're beginning to plan the next iteration of the **Getty Vocabularies** infrastructure:

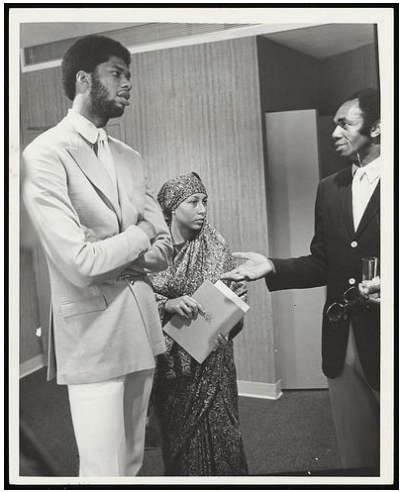
Working to understand how the multiple contexts of our **audiences** can be supported—and how new ways of working impact the platform.



And we're using the **platforms** and **standards** we've put in place to enable collaboration across the field:

Working with the Smithsonian to provide **joint access and discovery** for millions of images from the photo morgue of magazines such as Ebony and Jet.





Getty

Why do we do Linked Data?

The value is not in the technologies or the ontologies we use.

Why do we do Linked Data?

The value is in the **ecosystem**—information in varied context for different applications.

The value is in the **audience**—supporting user needs and conceptual models.

And it's in the **community**—allowing data and code to be used beyond the Getty.

Why do we do Linked Data?

We do it for humans.

Thank You.

Contact me at **dnewbury@getty.edu**

David Newbury
Assistant Director, Software and User Experience, Getty

CNI Spring Project Briefings, March 26, 2024

Getty